Characterizing internal models of the visual environment

Micha Engeser[1,2,3,+], Susan Ajith[1,4,5,+], Ilker Duymaz[1], Gongting Wang[1,6], Matthew J. Foxwell[7], Radoslaw M. Cichy[6], David Pitcher[7], Daniel Kaiser[1,3,*]

[1] *Department of Mathematics and Computer Science, Physics, Geography, Justus-Liebig-Universität Gießen, Germany*

[2] *Neural Circuits, Consciousness and Cognition Research Group, Max Planck Institute of Empirical Aesthetics, Frankfurt am Main, Germany*

[3] *Center for Mind, Brain and Behavior (CMBB), Philipps-Universität Marburg and Justus-Liebig-Universität Gießen, Germany*

[4] *Department of Medicine, Justus-Liebig-Universität Gießen, Germany*

[5] *Vision and Computational Cognition Group, Max Planck Institute for Human Cognitive and Brain Sciences, Germany*

[6] *Department of Education and Psychology, Freie Universität Berlin, Germany*

[7] *Department of Psychology, University of York, UK*

[+] *These authors contributed equally*

* *correspondence to:*

Prof. Dr. Daniel Kaiser
Department of Mathematics and Computer Science, Physics, Geography
Justus Liebig University Giessen
Arndtstr. 2
35392 Giessen
danielkaiser.net@gmail.com

Abstract

Despite the complexity of real-world environments, natural vision is seamlessly efficient. To explain this efficiency, researchers often use predictive processing frameworks, in which perceptual efficiency is determined by the match between the visual input and internal models of what the world should look like. In natural scene processing, predictions derived from our internal models of a scene should play a particularly important role, given the highly reliable statistical structure of our environment. Despite their importance for scene perception, we still do not fully understand what is contained in our internal models of the environment. Here, we argue that the current literature on scene perception disproportionately focuses on an experimental approach that tries to infer the contents of internal models from arbitrary, experimenter-driven manipulations in stimulus characteristics. To make progress, additional participant-driven approaches are needed: Rather than solely relying on manipulating the input to the visual system, researchers should adopt a complementary approach focusing on participants' descriptions of what they believe constitutes a typical scene. Such descriptions promise to capture the contents of internal models in more unconstrained ways on the level of individual participants. Critically, the descriptions of internal models can in turn be used to predict the efficiency of scene perception. We highlight recent studies on memory and perception using innovative methodologies like line drawings to characterize internal representations. These emerging methods show that it is now time to also study natural scene perception from a different angle – starting with a characterization of individual's expectations about the world.

1 Natural vision and internal models of the world

Perceptual efficiency is often understood through the lens of predictive processing (Clark, 2013; de Lange et al., 2018; Huang & Rao, 2011; Keller & Mrsic-Flogel, 2018). In this framework, visual inputs are routinely compared against internal models, which are based on our expectations of what the world should look like. In the processing of natural environments, internal models should play a particularly helpful role (Kayser et al., 2004; Mirza et al., 2016): Natural scenes are reliably structured, with a global structure that is stable across instances of a category and objects placed in statistically predictable locations (Bar, 2004; Kaiser et al., 2019a; Kaiser & Cichy, 2021; Oliva & Torralba, 2007; Vo et al., 2019, Vo, 2021). The reliable structure of natural scenes should give rise to rich internal models that capture what a specific scene (e.g., a kitchen) should typically look like.

The study of vision as an inverse inference problem has its origins in Helmholtz' idea of perception (1867). Within this framework, the perceptual system uses prior knowledge about the world, obtained through experience, to infer the causes of proximal stimulus patterns. In this view, internal models of the world, which contain this prior knowledge, thus are critical determinants for further efficient natural perception. This was later highlighted by schema theory, which postulated that inputs are referenced against internal models (schemata) that reflect the structure of the world (for instance the likely object arrangements found in a scene; Mandler, 1984; Minsky, 1974). This concept has influenced early research on human scene perception (Biederman, 1972; Biederman et al., 1982) and memory (Brewer & Treyens, 1981; Mandler & Parker, 1976). In contemporary research, this idea reverberates in the use of predictive processing frameworks for explaining how we perceive (Bar, 2009; de Lange et al., 2018; Peelen et al., 2024) and explore (Henderson, 2017) natural scenes, as well as how they are analyzed in the brain (Kaiser et al., 2019b; Muckli et al., 2015; Naselaris et al., 2009).

Together, the currently favored theoretical frameworks converge towards a view in which the contents of our internal models shape how we perceive the world around us. This assertion directly prompts critical questions: What exactly are the contents of the internal models that guide natural vision? Furthermore, given the variability in visual experiences, how do these models differ across individuals? Can these differences in internal models explain individual differences in visual perception? Here, we discuss how researchers in the domain of natural scene perception have attempted to characterize internal models during the last decades and delineate how the limitations of this approach necessitate complementary methods for addressing the questions raised above. Subsequently, we highlight an emerging trend of using creative methods suitable for making further progress toward a richer understanding of internal models across individuals. The central idea of the proposed methods is to prompt individual participants to report the characteristics of their internal model of a scene. We emphasize a selection of methods that could facilitate descriptions of the internal model, with a particular focus on the use of line drawings.

2 The classical approach to characterizing internal models

2.1 Probing internal models of scenes by manipulating input characteristics

Previous work was built on the assumption that the contents of internal models can be studied by varying the level of typicality (i.e., schema-congruence) of inputs to perceptual and cognitive systems. Researchers manipulated the typicality of scenes by altering various scene contents, such as the objects present in the scene or their spatial configurations. Such manipulations were usually at the level of objective priors, for example, repositioning an airplane in flight from the upper to the lower visual field (Kaiser & Cichy, 2018a) or creating semantic incongruencies such as placing a streetlight in an indoor and a living room lamp in an outdoor scene (Munneke et al., 2013). By doing so, they created diverse sets of stimuli that they considered in accordance or conflict with typical real-world experience, and thus with internal models (Fig. 1).

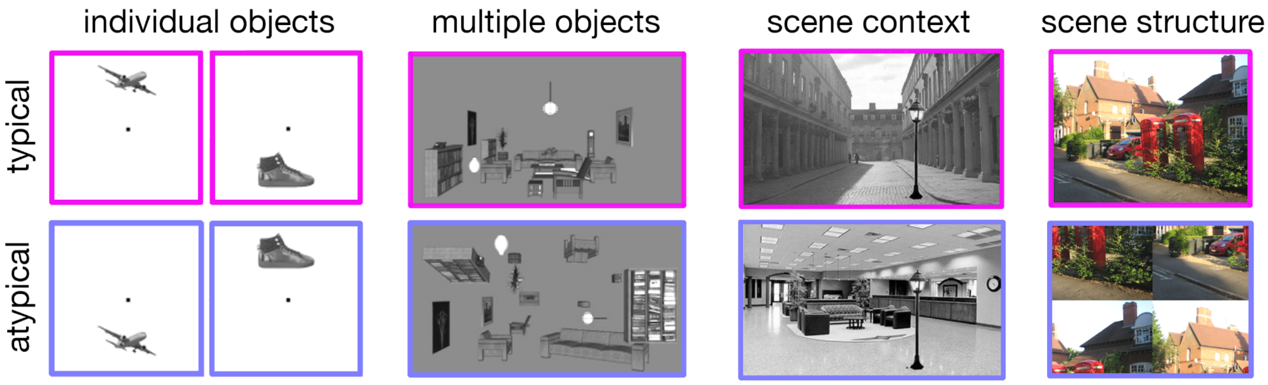

Figure 1. Understanding internal models through stimulus manipulation. *To understand which properties of natural environments are critical for internal models of the world, researchers have isolated and manipulated a set of regularities found in natural scenes. From left to right: manipulations in the typical positioning of individual objects across visual space (Kaiser & Cichy, 2018a), the typical composition of multiple objects across space (Figures reproduced from Bilalić et al., 2019), the semantic consistency between scenes and the objects they contain (Munneke et al.,*

2013), and the structural coherence of the scene (Kaiser et al., 2021). By comparing typically arranged stimuli with atypically arranged stimuli, they could show that the visual system preferentially processes stimuli in accordance with our priors about what the world should look like.

Through manipulating features that do or do not adhere to real-world structure, researchers were able to isolate various aspects of typical scene structure that facilitate more efficient perception and cortical processing. These include the positioning of individual objects across space (Kaiser et al., 2018; Kaiser & Cichy, 2018a, 2018b), spatial relationships between objects (Bilalić et al., 2019; Gronau et al., 2008; Gronau & Shachar, 2014; Kaiser et al., 2014; Kaiser & Peelen, 2018; Kim & Biederman, 2011; Roberts & Humphreys, 2010; Stein et al., 2015), contextual relationships between scenes and objects (Chen et al., 2022; Davenport & Potter, 2004; Faivre et al., 2019; Mudrik et al., 2010, 2011; Võ & Wolfe, 2013), and the spatial configuration of the scene as a whole (Biederman, 1972; Biederman et al., 1974; Kaiser et al., 2020a, 2020b). Together, these studies show that typical scene structure at multiple levels of description contributes to the efficient perception and neural representation of naturalistic inputs (for reviews, see: Bar, 2004; Castelhano & Krzyś, 2020; Kaiser et al., 2019a; Kaiser & Cichy, 2021; Oliva & Torralba, 2007; Peelen et al., 2024; Võ, 2021; Võ et al., 2019; Wolfe et al., 2011), suggesting that internal models contain rich information about the typical properties of natural scenes.

2.2 Challenges for the stimulus manipulation approach

While the approach of manipulating stimulus characteristics has led to significant advances in our understanding of the contents of our internal models of the world, this "classical" approach has several critical downsides.

First, it largely rests on the experimenter's intuition of what a typical scene looks like, and which factors construe its typicality. It cannot be taken for granted that researchers' intuitions cover those aspects of typical scenes reliably present across our real-world experience. To circumvent this problem, researchers have started using computational analyses to determine typical scene properties more objectively, for example by extracting object distributions across large scene databases (Boettcher et al., 2018; Bonner & Epstein, 2021; Gregorová et al., 2023; Kaiser et al., 2018, Kaiser, Quek, et al. 2019). These approaches can validate the relevant dimensions empirically, but they are still centered on the idea that the property selected by the researcher plays an important role. In natural vision, however, properties that are perceptually salient to observers can be relatively less important for our visual system, while others may be more intricate, but relatively more important.

Second, stimulus manipulation approaches only allow for independently manipulating particular stimulus features at a time, thus not capturing all possible interactions between them. Manipulating only a few stimulus dimensions while controlling for others is a key strategy in reductionist approaches to vision, which has proved particularly fruitful in distilling the response properties to simple visual stimuli across the visual cortex (Riesenhuber & Poggio, 1999). However, this approach inevitably reaches its limits when it comes to natural scenes (Felsen & Dan, 2005). Given their visual richness, scenes cannot be easily decomposed into a few orthogonal dimensions. What makes things even more challenging is that those different dimensions likely interact with each other: For instance, the types of objects appearing in a scene and their distribution across the scene depend on the scene's spatial geometry. In a kitchen scene, objects like an oven, stove, or sink are typically aligned along the walls, whereas smaller objects like utensils and cups are placed on horizontal surfaces like counters or tables. Studies only looking at object distributions or scene geometry separately may therefore miss critical interactions between these properties thus only providing limited insights into how strongly these dimensions influence vision in highly complex real-world environments. On a practical end, studies that look at such different factors tend to employ different experimental paradigms making it even harder to determine which factors contribute to what extent to efficient scene processing.

Third, many studies use artificial or unusual stimuli to create "atypical" scenes. For instance, in studies of scene-object congruence (Chen et al., 2022, Davenport, 2007; Davenport & Potter, 2004; Munneke et al., 2013; Öhlschläger & Võ, 2017), typical arrangements are easy to establish: a streetlight is typically found in outdoors, while living room lamp will be indoor. To study the effects of such congruence, researchers need to create additional incongruent conditions, in which the objects are positioned atypically: a living room lamp is shown on a street, and the streetlight in a living room. When comparing such conditions, one cannot be sure whether the difference can be attributed to enhanced processing of typically positioned objects or whether it arises from enhanced responses to the incongruence. On a cortical level, spatially separate regions code for congruent and incongruent conditions (Faivre et al., 2019), and recent behavioral work suggests that differences between congruent and incongruent conditions may indeed be driven by a "congruency cost", where the unexpected, incongruent objects gain a relative processing advantage (Spaak et al., 2022). In many studies with artificial stimuli, the problem is further aggravated by the nature of atypical conditions, which sometimes violate the laws of physics (Bilalić et al., 2019; Kaiser & Peelen, 2018), give rise to unusual inputs (Biederman, 1972; Biederman et al., 1974; Kaiser et al., 2020a), or introduce changes in visual image statistics (Underwood & Foulsham, 2006).

Finally, the stimulus manipulation approach neglects inter-individual differences in internal models. Most current research is based on the idea that there is a single "typical" scene that suffices for all observers. Consequently, pitting such typical scenes against clearly atypical scenes should yield reliable differences across participants, and thereby reveal critical properties of internal models. However, such an approach fails to account for inter-individual differences in internal models: although a typical kitchen will probably look fairly similar for two people, there may be critical differences, for instance in the identity and placement of objects across the scene. Such differences may be driven by the idiosyncratic visual diets that everyone experiences, but they may also link to cultural, linguistic, or socioeconomic factors (Barrett, 2020; Hartley, 2022). The stability of internal models across participants has allowed researchers to make progress with an approach that essentially ignores this variability. If we could instead harness this variability, we may find that there are characteristic differences in the way each of us perceives the world – based on our idiosyncratic priors of what the world looks like.

To overcome these critical differences and drive the field forward, an approach focused on getting descriptions directly from the participant bears enormous potential. Moving on, it is important to note that the following approaches constitute an addition to the existing toolkit for characterizing natural vision, and can be used complementarily to classical stimulus manipulation approaches.

3 A complementary approach for characterizing internal models of the world

3.1 Describing internal models

Here, we propose an approach to characterize individual participants' internal models of scenes in a much more unconstrained way, without any prior assumptions about their properties (Fig. 2). Instead of manipulating stimulus characteristics, and thereby the input to the visual system, we can ask participants to first provide "descriptions" about the contents of their internal model that can then be used to gain insights and inform further testing. By obtaining such descriptions of the internal model we can characterize what constitutes a typical scene for individual observers. For instance, while manipulating object-scene congruences possibly tells us about certain "objective" priors shared across individuals, subjective priors such as the contents of a typical bedroom can vary across individuals. With descriptive methods, such individually specific priors can be captured and we can start to understand how individuals converge and diverge in their conceptions of scene typicality. We can in turn use these insights to design smarter experiments that directly probe the most relevant dimensions of internal models of the world. Moreover, we can use our knowledge about the contents of these models to make targeted predictions about processing efficiency for individual scenes.

Importantly, this approach circumvents critical limitations of the classical stimulus manipulation approach. First, researchers do not need to use their intuitions about which features constrain internal models of natural scenes, as they can rather rely on the features emerging from participants' descriptions of typical scenes. Second, multiple interacting feature dimensions can be studied at once, as these will inherently be present in the descriptions. Third, there is no immediate need to artificially create atypical stimuli: through obtaining descriptors of the content of individual participants' internal models, the individual typicality of a range of stimuli can be defined based on their similarity to these descriptors. Finally, and critically, previous studies have overlooked inter-individual differences due to their inherent focus on understanding general mechanisms involved in vision. This approach pushes this boundary by increasing the explainable variance in the data and understanding processes that are both general and idiosyncratic in natural visual perception.

How can we experimentally obtain such descriptions of internal models? In the following, we will discuss methods that enable individual participants to report on how a typical exemplar of a particular scene should look like. One strong candidate is drawing, which has proven a versatile tool for transforming mental representations into visible descriptions with rich details (Fan et al., 2023). We also discuss other complementary methods including scene arrangement, verbal descriptions, and neuroimaging-based techniques. The studies we highlight are primarily focused on memory and perception, showcasing the potential of these methods for characterizing the contents of internal representations.

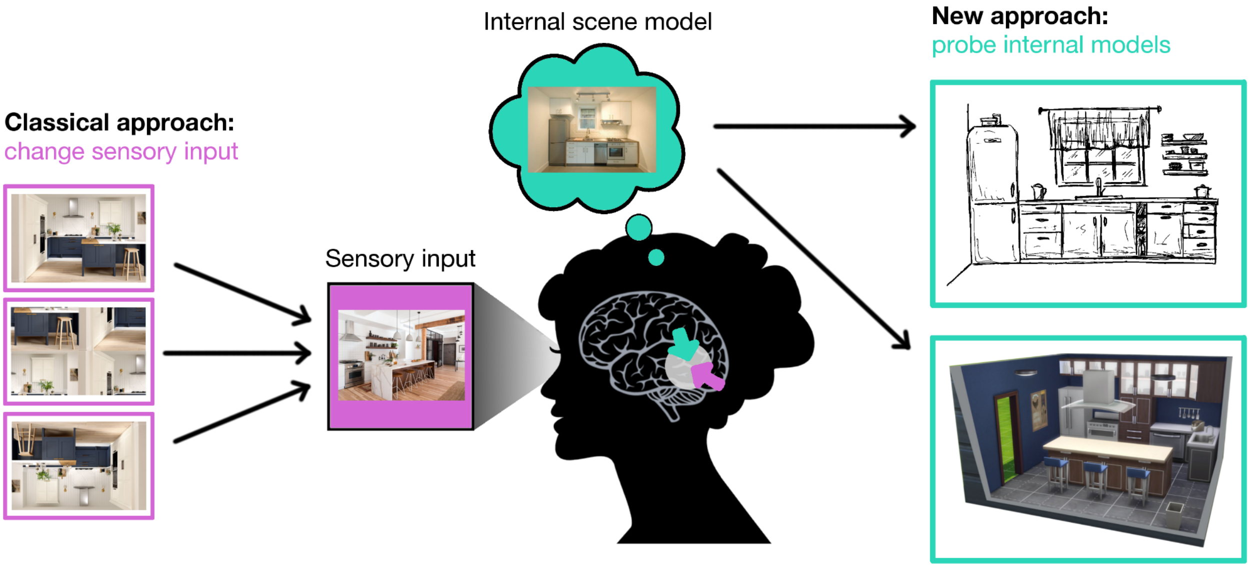

Figure 2. ***A complementary approach for studying internal models of the world****. The classical approach aims at discovering properties of internal models through stimulus manipulation, for instance by manipulating a scene's global structure. Here, we highlight a complementary approach, in which the contents of internal models are described by observers, for instance through line drawing (where people draw typical versions of scenes) or scene arrangement methods (where people arrange physical or virtual scenes in typical ways). These descriptions can in turn be used to derive targeted predictions about processing efficiency for a set of inputs.*

3.2 Line drawings as descriptors of internal models

Line drawings can be seen as functional abstractions of the ways in which we see the world in the sense that they "exploit the underlying neural codes of vision" (Sayim & Cavanagh, 2011): When we draw an object or a scene, our drawing tend to focus on what conveys the most essential details of visual images in a form that abstracts away from irrelevant detail. For instance, important boundaries between objects and surfaces are highlighted and will thereby reflect a parsing of the scene into behaviorally relevant units. Line drawings of scenes are recognized with virtually identical efficiency as scene photographs (Biederman & Ju, 1988), likely because they preserve critical information about the curvature and intersection of contours (Barrow & Tenenbaum, 1981; Walther et al., 2011). Further, neuroimaging work has shown that line drawings yield characteristic category-specific neural activation patterns in the high-level visual cortex (Singer et al., 2023; Walther et al., 2011). Beyond the theoretical considerations, they also offer practical advantages, such as being easy for participants to generate and enabling the creation of rich scenes in an (almost) unconstrained manner. These qualities render line drawings an ideal candidate for the description of internal models. This has recently been exploited in a set of studies that used line drawings to determine the contents of internal representations that are otherwise hard to access directly (Fan et al., 2023). In the following, we highlight how these advantages have been leveraged in clinical settings and developmental studies, before turning towards research in memory and perception.

3.2.1 Line drawings in clinical assessment

Drawings not only tell us about what constitutes typicality in mental representations but can also inform us about divergences from typicality. The use of drawings has a long history in clinical assessment to diagnose and classify visual impairments such as agnosia (Bauer, 2006), spatial neglect (Agrell & Dehlin, 1998), or different types of neurodegenerative disease (Cahn et al., 1996; Wechsler, 2009). Through drawings, such impairments can be quantified without requiring the cognitive processes needed for language-based tests. Furthermore, systematic differences in drawing tasks have been described for individuals with psychiatric disorders including autism spectrum disorder (Booth et al., 2003; Shi et al., 2021) and schizophrenia (Bozikas et al., 2004; Kaneda et al., 2010). These two cases are particularly interesting since it has been proposed that these conditions are associated with compromised predictive processing (Fletcher & Frith, 2009; Pellicano & Burr, 2012; Seymour et al., 2013; Van de Cruys et al., 2014). Alterations to drawings in these disorders may thus be linked to alterations in internal models of the world, as others have argued as well (Morgan et al., 2019).

3.2.2 Line drawings in developmental research

In a similar vein, drawings have proven extremely useful for tracking internal models across development, as they allow children to delineate the contents of internal models without the use of language. For instance, drawings have been used to assess the emergence of detail in visual representations of objects (Karmiloff-Smith, 1990; Long et al., 2019, 2024). In recent work, Long and colleagues (2019) asked a large group of children across different ages to draw various everyday objects, revealing characteristic changes in drawings across development (Fig. 3a). Further investigation revealed that the content of children's drawings predicted their ability to recognize visual objects (Long et al., 2024), suggesting a link between the mental representation revealed by drawings and the ability of the visual system to process critical object details. A similar link between visual production and recognition has also been found in neuroimaging work in adult participants (Fan et al., 2020): when adults practice visual production of categories, their neural representations of these categories become more distinctive. Together, these findings show that drawings mirror the ability of the visual system to efficiently perceive visual inputs.

3.2.3 Line drawings in studying memory

The use of drawings for studying internal representations in healthy adults has recently received renewed attention in the memory literature (see Fan et al., 2023 for a review). Here, drawings are useful because they allow researchers to study complex stimuli in free-recall paradigms, eliminating the need for recognition paradigms. Drawings thereby allow researchers to quantify fine-grained detail in participants' memory representations. For natural scenes, this method has recently been used by Bainbridge and colleagues (Bainbridge et al., 2019), who asked participants to freely recall scenes in a drawing paradigm and uncovered a surprising amount of detail in the representations of scenes and objects. These elements can be salvaged to study alterations in participants' memory as a function of stimulus characteristics. In another study, Bainbridge and colleagues (2021) probed object and scene representations for semantically congruent and incongruent scene-object combinations (Figure 3b). They could show that semantically inconsistent objects are remembered more vividly than semantically consistent objects – however, this came at the expense of weaker memory for other scene characteristics in the incongruent images. This result shows that drawings contain fine-grained information about internal representations held in memory.

Drawings are also used to determine the degree of boundary extension in natural scene images (Bainbridge & Baker, 2020), a phenomenon in which scenes drawn from memory tend to exhibit extended boundaries, compared to the originally memorized image (Intraub & Richardson, 1989). Using detailed analysis across a large stimulus set, Bainbridge & Baker (2020) showed that this effect ultimately is indeed stimulus-dependent: for some stimuli, boundaries are extended, and for others, they are compressed (Figure 4c). This effect is likely attributable to differences in the studied scenes' viewpoint and geometry, where the boundaries of "narrower" scenes tend to be subsequently extended, while for "wider" scenes they are compressed (Park et al., 2024). In this tradeoff, the depth of field plays a critical role, such that a naturalistic depth of field leads to a larger boundary extension than a non-naturalistic depth of field (Gandolfo et al., 2023). These results again showcase that drawings can be successfully employed to quantify key characteristics of internal representations.

3.2.4 Line drawings in visual perception

In vision science, drawings have been used more sparingly. In the predictive processing literature, perception is often described as a generative model in which a percept is constructed by employing the internal world model to infer the most likely cause of sensory input (Clark, 2013; Friston, 2005; Huang & Rao, 2011). In this framework, the brain generates hypotheses about how the world should look like. Drawings may offer researchers a tool to access these hypotheses by serving as an extension of the generative process into a visible format that allows us to infer the content and structure of the internal model from which these hypotheses were generated.

In line with this idea, drawings have been used to capture representations of inferred visual content in the visual cortex (Morgan et al., 2019). In this study, participants viewed natural scene images, in which one quadrant was occluded. The authors recorded fMRI activity from areas of the early visual cortex that exclusively respond to input from the occluded quadrant, and thus do not receive any direct visual information. They could show that there nonetheless was activity in the visual cortex region that remained unstimulated – the brain appears to "fill in" information from the surrounding context. Outside the scanner, they asked participants to draw what they thought should appear in the occluded quadrant. They then tested whether the activation patterns of different scenes could be predicted by the features of the differences in the drawings for the different scenes. They found that image-based low-level features of the drawings (for example their global spatial envelope, measured using a GIST model; Oliva & Torralba, 2001) predicted scene-specific activations in the unstimulated area of the primary visual cortex (Figure 4d). This result highlights how drawings can be used to map from behavior to neural representations during natural vision.

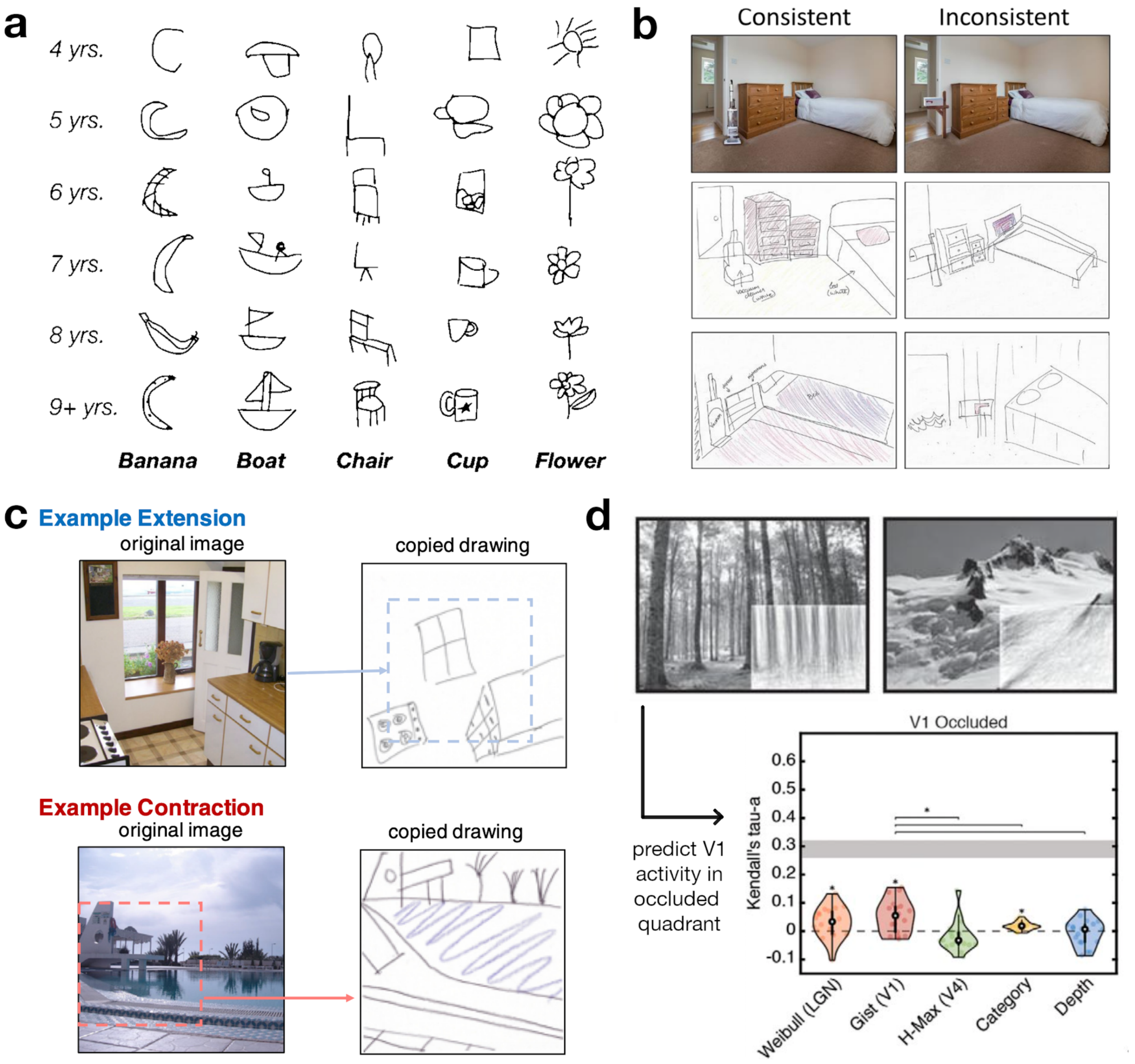

Figure 3. Using drawings to describe representations in development, memory, and perception. *a) In developmental research, the use of drawings allows researchers to gain insights into the emergence of detailed visual object representations (Long et al., 2019). These drawings can in turn be used to predict the maturation of object recognition across development (Figure reproduced from Long et al., 2024). b) In memory research, drawings can be used to quantify memory precision in free recall paradigms. Using drawings, Bainbridge et al. (2021) showed that relative to scenes with consistent objects, scenes with inconsistent objects are remembered with more detail about the consistent object, but with less detail about the scene context. (Figure reproduced from Bainbridge et al. 2021). c) Using a similar free recall paradigm, Bainbridge & Baker (2020) showed that scene boundaries are extended or compressed in memory, depending on the viewpoint and geometry of the original scene. d) In perception research, drawings can be used to probe how internal models facilitate the cortical filling-in of missing information. Participants' drawings of what should be present in an*

occluded quadrant predict neural activation: response patterns in areas of primary visual cortex (V1) that respond to the occluded quadrant are well explained by visual low-level visual features of these drawings (Morgan et al., 2019).

In a more recent study, Wang, Foxwell, and colleagues (2024) used drawings as a readout of individual participants' internal models of visual scenes (Figure 4). In this study, participants were asked to draw typical versions of a set of natural scene categories (e.g., kitchens or living rooms). These drawings were converted into standardized 3D renders to control for different drawing abilities and styles. If the drawings capture properties of individual participants' internal models for a scene category, then scenes that are more similar to these drawings should be perceived more efficiently. Indeed, participants were more accurate in categorizing renders that were constructed from their *own* drawings (and were thus more similar to their *own* internal models) than in categorizing renders based on *other* participants' drawings (which were more dissimilar to their *own* internal models). The authors further showed that the similarity to the scene renders based on participants' own drawings (measured by a deep neural network model) predicted categorization accuracy on other rendered scenes. This result demonstrates how drawings can be used to make personalized predictions about the efficiency of perception – derived from only a single drawing of a typical scene. Complementary EEG work (Wang et al., 2024) showed that neural representations of scenes that are similar to participants' drawings (and thus their internal models) are enhanced during perceptual processing unfolding in the initial 250ms of visual analysis. This suggests a rapid interaction between an individual's visual inputs and internal model during natural vision.

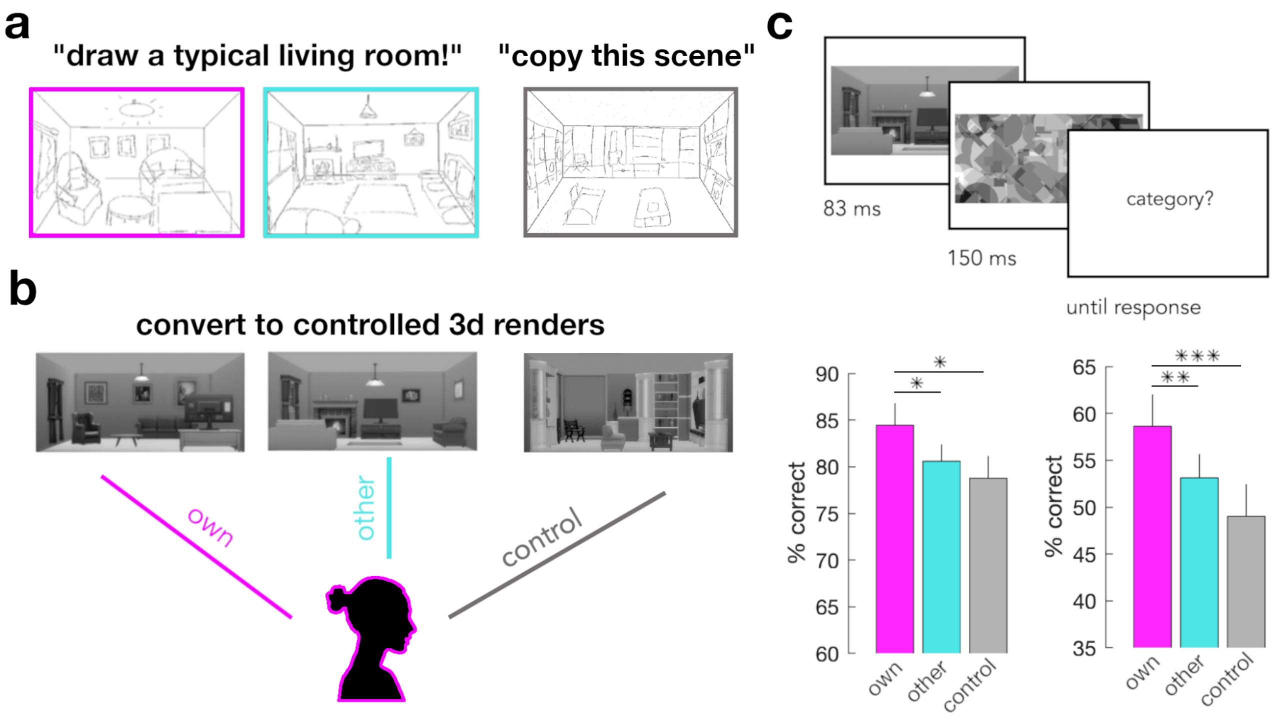

Figure 4. Using drawings to link individual differences in internal models to idiosyncrasies in perception. *a) To assess the contents of internal models for real-world scenes, participants drew typical versions of scene categories (here: living rooms). b) These drawings were converted to 3D renders to control for visual differences. c) During the subsequent categorization task, participants categorized briefly presented renders. Critically, they viewed renders based on their own drawings ("own" condition), other participants' drawings ("other" condition), or renders created from scenes participants previously copied from a photograph ("control" condition, designed to control for drawing-related familiarity effects). Participants more accurately categorized renders from the "own" condition than from the "other" or "control" conditions, suggesting that similarity to internal models on the individual level modulates scene processing in idiosyncratic ways. This result was replicated across two independent experiments with 2 (left) or 6 (right) scene categories (Wang, Foxwell, et al., 2024).*

However, the use of line drawings also brings about some limitations such as the challenge of quantifying the contents of drawings in objective ways and handling the substantial inter-subject variability in drawing abilities and style. Variation in drawing expertise has been associated with inter-individual differences in cognitive and perceptual abilities such as visual imagery, shape encoding, and detection, as well as the allocation of visual attention and working memory (Calabrese & Marucci, 2006; Chamberlain et al., 2019, 2021; Kozbelt, 2001; Perdreau & Cavanagh, 2015). Future research using drawings to access idiosyncrasies in internal representations needs to take these factors into account. Recently, different approaches to minimizing potentially confounding factors have been put forward: Carefully designing drawing experiments with sufficient sample sizes and suitable control conditions can mitigate variance related to drawing abilities. For example, Wang, Foxwell, and colleagues introduced a condition in which all participants copied the same scene photograph after drawing what they think would be the most typical version of that scene. Thereby they were able to control for the effects of drawing ability, familiarity, or memory in a subsequent categorization task based on these drawings. Further, innovative methods like pen-tracking, computer vision, or online crowdsourcing provide objective and reproducible tools for quantifying drawings (Bainbridge, 2022; Fan et al., 2023).

To sum up, drawings have proven to be a powerful tool for describing internal representations and have advanced our understanding of the precision of visual memory, the development of visual object representation, and the perception and neural representations of scenes. Nevertheless, the methodology of drawings comes with certain limitations, as discussed above. In the following, we briefly discuss three alternative methods for characterizing the contents of internal models, which offer complementary strengths to the limitations associated with drawing.

3.3 Alternative descriptors of internal models

Although drawings are a powerful tool for participants to provide rich descriptions of their internal model, various alternative techniques can provide descriptions while being less dependent on participants' drawings skills or fine-motor abilities in general. The first two approaches discussed in this section build on a very similar principle as drawings by enabling participants to describe their internal model using physical or virtual objects, as well as language. Finally, we highlight an approach that goes one step further by attempting to directly infer characteristics of subjects' internal models from measuring brain recordings without relying on overt reports.

3.3.1 Scene arrangement

In scene arrangement paradigms, participants create a scene by arranging a set of candidate objects provided by the experimenter. Although such methods often limit participants’ degrees of freedom in describing their internal models (e.g., because of fixed object exemplars available for arrangement), it mitigates the influence of inter-individual variability in drawings, which may not relate to the variability in internal models.

Scene arrangement tasks can be realized in real-world experiments that use physical objects. For example, Öhlschläger & Võ (2020) used a scene arrangement task to explore the emergence of structured scene representations across development. They asked their participants to arrange a set of miniature objects across a dollhouse (Figure 5a). Using this constructive measure, the authors could test how children across different age groups honour semantic relationships between objects (e.g., chairs and tables appear together in a dining room) or spatial regularities (e.g., chairs face the dining table) when equipping their doll houses. Results showed that children as young as 3 years respected semantic relationships but not spatial regularities among multiple related objects, while children over 4 years successfully arranged the objects across the dollhouse in semantically and spatially congruent ways. Using the same task, Bahn and colleagues (2025) could show that the performance on scene arrangement tasks somewhat covaries with the emergence of language during development, suggesting a link between the linguistic organization semantic concepts and the visual rules that structure natural scenes.

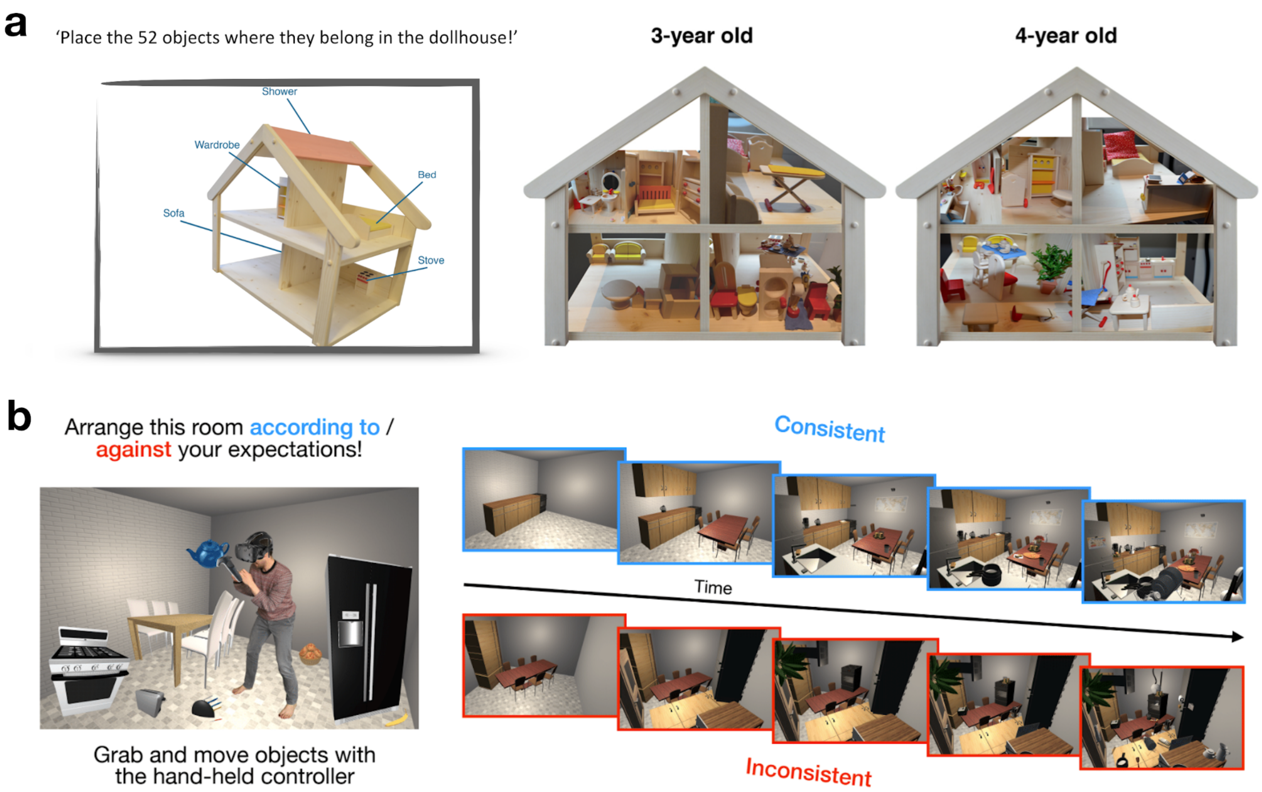
a
'Place the 52 objects where they belong in the dollhouse!'
Shower
Wardrobe
Bed
Sofa
Stove
3-year old
4-year old
b
Arrange this room according to / against your expectations!
Grab and move objects with the hand-held controller
Consistent
Time
Inconsistent

Figure 5. ***Using explicit scene arrangement to describe internal models.*** *(a) Children of different age groups were asked to arrange a set of miniature objects across a dollhouse. Object arrangements showed that children first appreciate semantic object similarities and only later incorporate the typical spatial organization across groups of objects* (Öhlschläger & Võ, 2020). *(b)*

Participants arranged objects in a VR environment into typical or atypical configurations. In subsequent search and memory tasks, participants performed better when the task was situated in the scenes they constructed in a typical fashion (Draschkow & Võ, 2017).

While such real-world scene arrangement tasks are intuitive for participants and offer a window into perception and action at the same time, they are relatively constrained: real objects need to be supplied, and they need to be moved around in physical space. To this end, emerging possibilities in virtual reality (VR) experiments allow for conducting similar studies with highly controlled, easily manipulable, and interactive environments while maintaining ecological validity (van den Oever et al., 2022; Wilson & Soranzo, 2015). Consequently, VR environments provide participants with a tool where they can arrange virtual worlds in accordance (or in disagreement) with their internal models of the real world. A noteworthy example is provided by a study by Draschkow and Võ (2017), who asked participants to construct scenes that concurred with their internal models of typical scenes (e.g., placing the objects in a bathroom in a typical fashion) or that violated their internal models (e.g., placing the same bathroom objects in an atypical fashion) (Figure 5b). Using this method, they could show that participants more successfully search and memorize scenes they arranged in typical ways, compared to the scenes they deliberately arranged in an atypical way. The results showcase that descriptors of internal models, as captured by explicit scene arrangement, can in turn be used to test perception and memory in environments that are specifically tailored to individual participants' models of the world. The use of VR environments further allows researchers to devise creative visual learning experiments in virtual worlds with changed scene statistics, as well as (active) visual search paradigms with overtly visible (Beitner et al., 2021; David et al., 2021; Schuetz et al., 2024) or hidden (David & Võ, 2022) objects.

The scene arrangement method provides an interactive environment where typical object configurations can be constructed intuitively. A physical arrangement of objects is less dependent on expertise than drawings making it more suitable for comparisons across the development of children, however, it is limited to a definite number of objects reducing the richness of the report. This can be partially mitigated by the use of VR, where a much larger variability of objects could be made available. Yet, such environments can be more complicated to use (for the participants as well as the researchers), potentially impairing accessibility for example for children, elderly people, or certain clinical populations.

3.3.2 Verbal descriptions

Focusing more directly on conceptual – rather than visual – attributes of scenes, verbal descriptions of scenes offer another effective tool for characterizing internal representations of scenes. For instance, Greene et al. (2015) used verbal descriptions to study how prior experiences affect perception. Participants were asked to describe probable (i.e., typically encountered) and improbable scenes presented for varying durations. Results showed that the quality of descriptions of improbable scenes deteriorated faster with shorter presentation duration and more often included objects not present in the images, compared to probable scenes. This suggests that people perceive typical scenes more quickly and accurately, underscoring the utility of language for providing rich scene characterizations.

Another recent study used verbal descriptions to capture individual differences in scene perception (Kollenda et al., 2024). Here, verbal scene descriptions of individual participants were used to study idiosyncrasies in their gaze behavior when exploring the same scenes visually. Participants freely viewed a set of natural scenes and subsequently provided verbal descriptions of the same scenes. Pairwise inter-subject similarities in fixation patterns between observers could be predicted by the inter-subject similarities in scene descriptions, particularly in the use of nouns: Participants who mentioned people more often in their descriptions also looked at people more prominently during exploration, and participants who mentioned text more often spent more time looking at text. This finding highlights the potential of verbal descriptors to capture information about scene representations on the individual level, which can be utilized to make predictions about how we explore or perceive natural scenes.

Future studies could employ verbal descriptions to gauge the contents of internal models directly, revealing the conceptual factors that organize our priors. Specifically, similar to the paradigm used by Wang, Foxwell, and colleagues (2024) wherein participants were asked to draw typical versions of a set of natural scene categories and then tested on categorization for scenes similar or dissimilar to these drawings, participants could alternatively provide verbal descriptions what they think a typical exemplar of a specific scene category should look like. The correspondence of any given scene image with the verbal descriptions provided by individual participants could in turn be used to predict perceptual efficiency or neural responses for these scenes on the individual level. Alternatively, generative text-to-image models could be used to generate stimulus materials that are in accordance with individual participants' verbal descriptions or deviate from them in targeted ways (by manipulating the descriptions supplied to the model).

Verbal reports are comparably easy to obtain for most subject populations, however, they are susceptible to language abilities. Verbal descriptions can carry a lot of detail but may be less precise and potentially sparser than drawings. For example, a participant might say something like “The room has a table in the right corner”. Yet, a drawing of this table conveys many details about what kind of table or how it is placed in the corner that are difficult to express using words.

3.3.3 Neural quantification of internal models

The methods for describing internal models discussed thus far rely on behavioral reports. Alternatively, researchers could read out the content of participants’ internal models from neural responses. Such direct read-out from the brain would be an exciting future prospect because brain activity recorded during simple visual tasks or passive fixation is potentially less prone to subjective biases and task-specific demand characteristics. A few recent studies that suggest such a potential are discussed below.

The response of the human visual system to scenes is influenced by their predictability. Scenes (and objects) that align with internal models more strongly produce more diagnostic, “sharpened” neural responses: For instance, Torralbo and colleagues (2013) demonstrated that atypical scenes elicited stronger brain responses, while more typical exemplars revealed higher decodability (i.e., better discrimination between scene categories) in scene-selective brain areas. A similar pattern was observed for objects, where more typical objects are associated with weaker univariate signals but more pronounced category information in object-selective or semantic brain regions (Clarke & Tyler, 2014; Delhaye et al., 2023; Fairhall & Caramazza, 2013; Martin et al., 2018; Santi et al., 2016).

This variation of visual responses can be used to derive brain-based measures of stimulus typicality. For example, Iordan and colleagues (2016) generated a brain activity prototype for each category by averaging voxel activity patterns for all exemplars of a stimulus category. They found that neural responses to exemplars rated as typical were more similar to the prototype than neural responses to atypical objects. Similarly, Davis and Poldrack (2014) quantified neural typicality by comparing an exemplar’s activity pattern to all other members of that category, following the premise that a typical exemplar shares more attributes with other members of its category than atypical exemplars (Rosch & Mervis, 1975). They found that a more central position in this representational space was linked to higher typicality ratings.

These studies show how the brain responds differently to typical and atypical scenes, and how we can employ these systematic differences to predict the typicality of a scene from neuroimaging data. Developing models of neural typicality for individual subjects could provide a data-driven technique to determine a subject's internal model of a scene. Thereby, we could get closer to understanding the neural correlates of individual differences in the processing of visual information (Lin & Lau, 2024). However, more work is needed to characterize how typicality is processed in the brain before purely neuroimaging-based methods can be used to read out a person's internal model.

4 Challenges in describing internal models

The methods described above enable researchers to infer the content of internal models with fewer constraints and prior assumptions about their properties than classical approaches. We underscored the virtues of these methods for providing new insights into how internal models shape perception on the group level and for an individual. However, there are a series of challenges to this approach. We will highlight three of these challenges below, and outline what we gain from solving them.

First, behavioral descriptions of internal models, such as drawings or verbal reports, are subjective reports. Can we assume that such introspective insights are reliable? Introspection is often disregarded as inherently problematic because observers may not be able to reliably characterize their internal representations as the introspective process itself invariably taints them (Engelbert & Carruthers, 2010; Schwitzgebel, 2008). However, this view has been challenged, most prominently by Gestalt psychologists (Koffka, 1924), but also more recently (Jack & Roepstorff, 2002; Jack & Shallice, 2001; Locke, 2009), with proponents arguing that introspection offers converging and additional information compared to analytic approaches. In the case of internal scene representations, the quality of introspective insight can be addressed empirically. If we take descriptions (i.e., obtained through drawings) at face value, we may indeed primarily measure subjective interpretations of internal states. If, however, we use these descriptions to inform experimental design and stimulus manipulations, we will be able to quantify whether introspective insights about internal models can be used to predict the efficiency of information processing in perceptual and cognitive systems. In the future, we thus need to combine stimulus manipulation approaches with approaches for describing internal models. Our review therefore does not make the case to replace or overcome classical approaches to studying scene vision – we rather advocate for adding a complementary method to the available toolkit for characterizing natural vision.

Second, some of the outlined methods have the additional challenge of separating informative differences from incidental variance across individuals. This is particularly relevant for drawings, where different drawing styles and abilities (Chamberlain, 2018; Chamberlain et al., 2014) introduce substantial variation that is not directly related to the content of internal representations as discussed above. Verbal descriptions or the scene arrangement method may also be compromised by expertise in language or digital skills, respectively. Moreover, methods like scene arrangement or verbal reports may bias the provided descriptions by offering a limited number of available objects or words that can be used to express an idea. A careful choice of methods, control conditions, sample size, and analysis approach is required to minimize these shortcomings and unfold the complementary strengths of these techniques.

Third, internal models are likely to dynamically change across time, behavioral goals, and mental states. Yet, most of the methods highlighted here yield single descriptions of participants' internal models. This implies that there is a single internal model for a given scene category — similar to a single "attractor point" in multidimensional space. However, the reality is likely more complex, and internal models may encompass a range of multiple typical scene configurations (e.g., bathroom templates for a bathroom in a private home versus a public bathroom). Moreover, internal models and their interaction with perception can be shaped by the context (e.g., in the form of precision weighting, Clark, 2013; Hohwy, 2020) and are sensitive to behavioral goals (Bracci & op de Beeck, 2023). This requires experiments that repeatedly quantify the contents of internal models within the same participants, thereby characterizing which aspects of the internal model remain stable and which aspects flexibly adapt to context.

5 Conclusion

We showed that classical approaches to characterizing people's internal models of naturalistic scenes through stimulus manipulation have notable limitations including an overreliance on a-priori assumptions about scene typicality, restricted possibilities for stimulus manipulation, the creation of highly artificial stimuli, and insensitivity to inter-individual differences. We therefore highlight a complementary methodological framework that allows for more unconstrained descriptors of participants' internal models. One promising method is the use of line drawings, which has recently opened new avenues in the study of visual memory and perception. Overall, we believe that natural vision research greatly benefits from methods with fewer constraints and prior assumptions about the nature of internal models, complementing traditional approaches. Embracing this approach could yield novel insights into how internal models differently shape perception across individuals and across cultural and linguistic contexts, and how alterations of internal models drive changes in visual processing across the lifespan and from health to disease.

6 Acknowledgements

This work was supported by an ERC Starting Grant (PEP, ERC-2022-STG 101076057; awarded to D.K.). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. We thank Melissa Võ for providing materials for Figure 4. The authors declare that there are no competing interests.

7 References

Agrell, B., & Dehlin, O. (1998). The clock-drawing test. *Age and Ageing*, *27*(3), 399–403. https://doi.org/10.1093/ageing/27.3.399

Bahn, D., Türk, D. D., Tsenkova, N., Schwarzer, G., Võ, M. L.-H., & Kauschke, C. (2025). Processing of Scene-Grammar Inconsistencies in Children with Developmental Language Disorder—Insights from Implicit and Explicit Measures. Brain Sciences, 15(2), Article 2. https://doi.org/10.3390/brainsci15020139

Bainbridge, W. A. (2022). A tutorial on capturing mental representations through drawing and crowd-sourced scoring. Behavior Research Methods, 54(2), 663–675. https://doi.org/10.3758/s13428-021-01672-9

Bainbridge, W. A., & Baker, C. I. (2020). Boundaries Extend and Contract in Scene Memory Depending on Image Properties. *Current Biology*, *30*(3), 537-543.e3. https://doi.org/10.1016/j.cub.2019.12.004

Bainbridge, W. A., Hall, E. H., & Baker, C. I. (2019). Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-018-07830-6

Bainbridge, W. A., Kwok, W. Y., & Baker, C. I. (2021). Disrupted object-scene semantics boost scene recall but diminish object recall in drawings from memory. *Memory & Cognition*, *49*(8), 1568–1582. https://doi.org/10.3758/s13421-021-01180-3

Bar, M. (2004). Visual objects in context. Nature Reviews Neuroscience, 5(8), 617–629. https://doi.org/10.1038/nrn1476

Bar, M. (2009). The proactive brain: Memory for predictions. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, *364*(1521), 1235–1243. https://doi.org/10.1098/rstb.2008.0310

Barrett, H. C. (2020). Towards a Cognitive Science of the Human: Cross-Cultural Approaches and Their Urgency. *Trends in Cognitive Sciences*, *24*(8), 620–638. https://doi.org/10.1016/j.tics.2020.05.007

Barrow, H. G., & Tenenbaum, J. M. (1981). Interpreting line drawings as three-dimensional surfaces. *Artificial Intelligence*, *17*(1), 75–116. https://doi.org/10.1016/0004-3702(81)90021-7

Bauer, R. M. (2006). The Agnosias. In *Clinical neuropsychology: A pocket handbook for assessment, 2nd ed* (S. 508–533). American Psychological Association. https://doi.org/10.1037/11299-020

Beitner, J., Helbing, J., Draschkow, D., & Võ, M. L.-H. (2021). Get Your Guidance Going: Investigating the Activation of Spatial Priors for Efficient Search in Virtual Reality.

Brain Sciences, *11*(1), Article 1. https://doi.org/10.3390/brainsci11010044

Biederman, I. (1972). Perceiving Real-World Scenes. *Science*, *177*(4043), 77–80.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, *20*(1), 38–64. https://doi.org/10.1016/0010-0285(88)90024-2

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*(2), 143–177. https://doi.org/10.1016/0010-0285(82)90007-X

Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology*, *103*(3), 597–600. https://doi.org/10.1037/h0037158

Bilalić, M., Lindig, T., & Turella, L. (2019). Parsing rooms: The role of the PPA and RSC in perceiving object relations and spatial layout. *Brain Structure and Function*, *224*, 2505–2524.

Bracci, S., & op de Beeck, H. P. (2023). Understanding Human Object Vision: A Picture Is Worth a Thousand Representations. Annual Review of Psychology, 74(Volume 74, 2023), 113–135. https://doi.org/10.1146/annurev-psych-032720-041031

Boettcher, S. E. P., Draschkow, D., Dienhart, E., & Võ, M. L.-H. (2018). Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search. *Journal of Vision*, *18*(13), 11. https://doi.org/10.1167/18.13.11

Bonner, M. F., & Epstein, R. A. (2021). Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-24368-2

Booth, R., Charlton, R., Hughes, C., & Happé, F. (2003). Disentangling weak coherence and executive dysfunction: Planning drawing in autism and attention–deficit/hyperactivity disorder. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1430), 387–392. https://doi.org/10.1098/rstb.2002.1204

Bozikas, V. P., Kosmidis, M. H., Gamvrula, K., Hatzigeorgiadou, M., Kourtis, A., & Karavatos, A. (2004). Clock Drawing Test in patients with schizophrenia. *Psychiatry Research*, *121*(3), 229–238. https://doi.org/10.1016/j.psychres.2003.07.003

Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology*, *13*(2), 207–230. https://doi.org/10.1016/0010-0285(81)90008-6

Cahn, D. A., Salmon, D. P., Monsch, A. U., Butters, N., Wiederholt, W. C., Corey-Bloom, J., & Barrett-Connor, E. (1996). Screening for dementia of the alzheimer type in the community: The utility of the Clock Drawing Test. *Archives of Clinical*

Neuropsychology: The Official Journal of the National Academy of Neuropsychologists, *11*(6), 529–539.

Calabrese, L., & Marucci, F. S. (2006). The influence of expertise level on the visuo-spatial ability: Differences between experts and novices in imagery and drawing abilities. *Cognitive Processing*, *7*(1), 118–120. https://doi.org/10.1007/s10339-006-0094-2

Castelhano, M. S., & Krzyś, K. (2020). Rethinking Space: A Review of Perception, Attention, and Memory in Scene Processing. *Annual Review of Vision Science*, *6*, 563–586. https://doi.org/10.1146/annurev-vision-121219-081745

Chamberlain, R. (2018). Drawing as a Window Onto Expertise. *Current Directions in Psychological Science*, *27*(6), 501–507. https://doi.org/10.1177/0963721418797301

Chamberlain, R., Drake, J. E., Kozbelt, A., Hickman, R., Siev, J., & Wagemans, J. (2019). Artists as experts in visual cognition: An update. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(1), 58–73. https://doi.org/10.1037/aca0000156

Chamberlain, R., Kozbelt, A., Drake, J. E., & Wagemans, J. (2021). Learning to see by learning to draw: A longitudinal analysis of the relationship between representational drawing training and visuospatial skill. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(1), 76–90. https://doi.org/10.1037/aca0000243

Chamberlain, R., McManus, C., Riley, H., Rankin, Q., & Brunswick, N. (2014). Cain's house task revisited and revived: Extending theory and methodology for quantifying drawing accuracy. *Psychology of Aesthetics, Creativity, and the Arts*, *8*(2), 152–167. https://doi.org/10.1037/a0035635

Chen, L., Cichy, R. M., & Kaiser, D. (2022). Semantic scene-object consistency modulates N300/400 EEG components, but does not automatically facilitate object representations. *Cerebral Cortex*, *32*(16), 3553–3567. https://doi.org/10.1093/cercor/bhab433

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Clarke, A., & Tyler, L. K. (2014). Object-Specific Semantic Coding in Human Perirhinal Cortex. *Journal of Neuroscience*, *34*(14), 4766–4775. https://doi.org/10.1523/JNEUROSCI.2828-13.2014

Davenport, J. L. (2007). Consistency effects between objects in scenes. *Memory & Cognition*, *35*(3), 393–401. https://doi.org/10.3758/BF03193280

Davenport, J. L., & Potter, M. C. (2004). Scene Consistency in Object and Background Perception. *Psychological Science*, *15*(8), 559–564. https://doi.org/10.1111/j.0956-7976.2004.00719.x

David, E., Beitner, J., & Võ, M. L.-H. (2021). The importance of peripheral vision when searching 3D real-world scenes: A gaze-contingent study in virtual reality. *Journal of Vision*, *21*(7), 3. https://doi.org/10.1167/jov.21.7.3

David, E., & Võ, M. L.-H. (2022). Searching for hidden objects in 3D environments. *Journal of Vision*, *22*(14), 3901. https://doi.org/10.1167/jov.22.14.3901

Davis, T., & Poldrack, R. A. (2014). Quantifying the Internal Structure of Categories Using a Neural Typicality Measure. *Cerebral Cortex*, *24*(7), 1720–1737. https://doi.org/10.1093/cercor/bht014

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How Do Expectations Shape Perception? *Trends in Cognitive Sciences*, *22*(9), 764–779. https://doi.org/10.1016/j.tics.2018.06.002

Delhaye, E., Coco, M. I., Bahri, M. A., & Raposo, A. (2023). Typicality in the brain during semantic and episodic memory decisions. *Neuropsychologia*, *184*, 108529. https://doi.org/10.1016/j.neuropsychologia.2023.108529

Draschkow, D., & Võ, M. L.-H. (2017). Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific Reports*, *7*(1), 16471. https://doi.org/10.1038/s41598-017-16739-x. Copyrights license.

Engelbert, M., & Carruthers, P. (2010). Introspection. *WIREs Cognitive Science*, *1*(2), 245–253. https://doi.org/10.1002/wcs.4

Fairhall, S. L., & Caramazza, A. (2013). Brain Regions That Represent Amodal Conceptual Knowledge. *Journal of Neuroscience*, *33*(25), 10552–10558. https://doi.org/10.1523/JNEUROSCI.0051-13.2013

Faivre, N., Dubois, J., Schwartz, N., & Mudrik, L. (2019). Imaging object-scene relations processing in visible and invisible natural scenes. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-38654-z

Fan, J. E., Bainbridge, W. A., Chamberlain, R., & Wammes, J. D. (2023). Drawing as a versatile cognitive tool. *Nature Reviews Psychology*, *2*(9), 556–568. https://doi.org/10.1038/s44159-023-00212-w

Fan, J. E., Wammes, J. D., Gunn, J. B., Yamins, D. L. K., Norman, K. A., & Turk-Browne, N. B. (2020). Relating Visual Production and Recognition of Objects in Human Visual Cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *40*(8), 1710–1721. https://doi.org/10.1523/JNEUROSCI.1843-19.2019

Felsen, G., & Dan, Y. (2005). A natural approach to studying vision. *Nature Neuroscience*, *8*(12), 1643–1646. https://doi.org/10.1038/nn1608

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10*(1), 48–58. https://doi.org/10.1038/nrn2536

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Gandolfo, M., Nägele, H., & Peelen, M. V. (2023). Predictive Processing of Scene Layout Depends on Naturalistic Depth of Field. *Psychological Science*, *34*(3), 394–405. https://doi.org/10.1177/09567976221140341

Greene, M. R., Botros, A. P., Beck, D. M., & Fei-Fei, L. (2015). What you see is what you expect: Rapid scene understanding benefits from prior experience. *Attention, Perception, & Psychophysics*, *77*(4), 1239–1251. https://doi.org/10.3758/s13414-015-0859-8

Gregorová, K., Turini, J., Gagl, B., & Võ, M. L.-H. (2023). Access to meaning from visual input: Object and word frequency effects in categorization behavior. *Journal of Experimental Psychology: General*, *152*(10), 2861–2881. https://doi.org/10.1037/xge0001342

Gronau, N., Neta, M., & Bar, M. (2008). Integrated contextual representation for objects' identities and their locations. *Journal of Cognitive Neuroscience*, *20*(3), 371–388. https://doi.org/10.1162/jocn.2008.20027

Gronau, N., & Shachar, M. (2014). Contextual integration of visual objects necessitates attention. *Attention, Perception & Psychophysics*, *76*(3), 695–714. https://doi.org/10.3758/s13414-013-0617-8

Hartley, C. A. (2022). How do natural environments shape adaptive cognition across the lifespan? *Trends in Cognitive Sciences*, *26*(12), 1029–1030. https://doi.org/10.1016/j.tics.2022.10.002

Henderson, J. M. (2017). Gaze Control as Prediction. *Trends in Cognitive Sciences*, *21*(1), 15–23. https://doi.org/10.1016/j.tics.2016.11.003

Huang, Y., & Rao, R. P. N. (2011). Predictive coding. *WIREs Cognitive Science*, *2*(5), 580–593. https://doi.org/10.1002/wcs.142

Intraub, H., & Richardson, M. (1989). Wide-angle memories of close-up scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(2), 179–187. https://doi.org/10.1037/0278-7393.15.2.179

Iordan, M. C., Greene, M. R., Beck, D. M., & Fei-Fei, L. (2016). Typicality sharpens category representations in object-selective cortex. *NeuroImage*, *134*, 170–179. https://doi.org/10.1016/j.neuroimage.2016.04.012

Jack, A. I., & Roepstorff, A. (2002). Introspection and cognitive brain mapping: From stimulus-response to script-report. *Trends in Cognitive Sciences*, *6*(8), 333–339.

https://doi.org/10.1016/s1364-6613(02)01941-1

Jack, A. I., & Shallice, T. (2001). Introspective physicalism as an approach to the science of consciousness. *Cognition*, *79*(1), 161–196. https://doi.org/10.1016/S0010-0277(00)00128-1

Kaiser, D., & Cichy, R. M. (2018a). Typical visual-field locations enhance processing in object-selective channels of human occipital cortex. *Journal of Neurophysiology*, *120*(2), 848–853. https://doi.org/10.1152/jn.00229.2018

Kaiser, D., & Cichy, R. M. (2018b). Typical visual-field locations facilitate access to awareness for everyday objects. *Cognition*, *180*, 118–122. https://doi.org/10.1016/j.cognition.2018.07.009

Kaiser, D., & Cichy, R. M. (2021). Parts and Wholes in Scene Processing. *Journal of Cognitive Neuroscience*, *34*(1), 4–15. https://doi.org/10.1162/jocn_a_01788

Kaiser, D., Häberle, G., & Cichy, R. M. (2020a). Cortical sensitivity to natural scene structure. *Human Brain Mapping*, *41*(5), 1286–1295. https://doi.org/10.1002/hbm.24875

Kaiser, D., Häberle, G., & Cichy, R. M. (2020b). Real-world structure facilitates the rapid emergence of scene category information in visual brain signals. *Journal of Neurophysiology*, *124*(1), 145–151. https://doi.org/10.1152/jn.00164.2020

Kaiser, D., Häberle, G., & Cichy, R. M. (2021). Coherent natural scene structure facilitates the extraction of task-relevant object information in visual cortex. *NeuroImage*, *240*, 118365. https://doi.org/10.1016/j.neuroimage.2021.118365

Kaiser, D., Moeskops, M. M., & Cichy, R. M. (2018). Typical retinotopic locations impact the time course of object coding. NeuroImage, 176, 372–379. https://doi.org/10.1016/j.neuroimage.2018.05.006

Kaiser, D., & Peelen, M. V. (2018). Transformation from independent to integrative coding of multi-object arrangements in human visual cortex. NeuroImage, 169, 334–341. https://doi.org/10.1016/j.neuroimage.2017.12.065

Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object Vision in a Structured World. *Trends in Cognitive Sciences*, *23*(8), 672–685. https://doi.org/10.1016/j.tics.2019.04.013

Kaiser, D., Stein, T., & Peelen, M. V. (2014). Object grouping based on real-world regularities facilitates perception by reducing competitive interactions in visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(30), 11217–11222. https://doi.org/10.1073/pnas.1400559111

Kaneda, A., Yasui-Furukori, N., Saito, M., Sugawara, N., Nakagami, T., Furukori, H., & Kaneko, S. (2010). Characteristics of the tree-drawing test in chronic schizophrenia.

Psychiatry and Clinical Neurosciences, *64*(2), 141–148. https://doi.org/10.1111/j.1440-1819.2010.02071.x

Karmiloff-Smith, A. (1990). Constraints on representational change: Evidence from children's drawing. *Cognition*, *34*(1), 57–83. https://doi.org/10.1016/0010-0277(90)90031-E

Kayser, C., Körding, K. P., & König, P. (2004). Processing of complex stimuli and natural scenes in the visual cortex. *Current Opinion in Neurobiology*, *14*(4), 468–473. https://doi.org/10.1016/j.conb.2004.06.002

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive Processing: A Canonical Cortical Computation. *Neuron*, *100*(2), 424–435. https://doi.org/10.1016/j.neuron.2018.10.003

Kim, J. G., & Biederman, I. (2011). Where Do Objects Become Scenes? *Cerebral Cortex*, *21*(8), 1738–1746. https://doi.org/10.1093/cercor/bhq240

Koffka, K. (1924). Introspection and the Method of Psychology. *British Journal of Psychology*, *15*, 149–161.

Kollenda, D., Reher, A.-S. V., & de Haas, B. (2024). *Individual gaze predicts individual scene descriptions*. https://osf.io/nx7jy/download

Kozbelt, A. (2001). Artists as experts in visual cognition. *Visual Cognition*, *8*(6), 705–723. https://doi.org/10.1080/13506280042000090

Lin, Q., & Lau, H. (2024). *Individual differences in prefrontal coding of visual features* (S. 2024.05.09.588948). bioRxiv. https://doi.org/10.1101/2024.05.09.588948

Locke, E. A. (2009). It's Time We Brought Introspection Out of the Closet. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *4*(1), 24–25. https://doi.org/10.1111/j.1745-6924.2009.01090.x

Long, B., Fan, J., Chai, Z., & & Frank, M. C. (2019). Developmental changes in the ability to draw distinctive features of object categories. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*. https://par.nsf.gov/biblio/10128364-developmental-changes-ability-draw-distinctive-features-object-categories

Long, B., Fan, J. E., Huey, H., Chai, Z., & Frank, M. C. (2024). Parallel developmental changes in children's production and recognition of line drawings of visual concepts. *Nature Communications*, *15*(1), Article 1. https://doi.org/10.1038/s41467-023-44529-9. Copyrights license.

Mandler, J. M. (1984). *Stories, Scripts, and Scenes: Aspects of Schema Theory*. Psychology Press. https://doi.org/10.4324/9781315802459

Mandler, J. M., & Parker, R. E. (1976). Memory for descriptive and spatial information in complex pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(1), 38–48. https://doi.org/10.1037/0278-7393.2.1.38

Martin, C. B., Douglas, D., Newsome, R. N., Man, L. L., & Barense, M. D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *eLife*, *7*, e31873. https://doi.org/10.7554/eLife.31873

Minsky, M. (1974). *A framework for representing knowledge*. MIT, Cambridge. https://dspace.mit.edu/bitstream/handle/1721.1/6089/AIM-306.pdf?%2520sequence%3D2

Mirza, M. B., Adams, R. A., Mathys, C. D., & Friston, K. J. (2016). Scene Construction, Visual Foraging, and Active Inference. *Frontiers in Computational Neuroscience*. https://doi.org/10.3389/fncom.2016.00056

Morgan, A. T., Petro, L. S., & Muckli, L. (2019). Scene Representations Conveyed by Cortical Feedback to Early Visual Cortex Can Be Described by Line Drawings. *Journal of Neuroscience*, *39*(47), 9410–9423. https://doi.org/10.1523/JNEUROSCI.0852-19.2019

Muckli, L., De Martino, F., Vizioli, L., Petro, L. S., Smith, F. W., Ugurbil, K., Goebel, R., & Yacoub, E. (2015). Contextual Feedback to Superficial Layers of V1. *Current Biology*, *25*(20), 2690–2695. https://doi.org/10.1016/j.cub.2015.08.057

Mudrik, L., Breska, A., Lamy, D., & Deouell, L. Y. (2011). Integration Without Awareness: Expanding the Limits of Unconscious Processing. *Psychological Science*, *22*(6), 764–770. https://doi.org/10.1177/0956797611408736

Mudrik, L., Lamy, D., & Deouell, L. Y. (2010). ERP evidence for context congruity effects during simultaneous object–scene processing. *Neuropsychologia*, *48*(2), 507–517. https://doi.org/10.1016/j.neuropsychologia.2009.10.011

Munneke, J., Brentari, V., & Peelen, M. (2013). The influence of scene context on object recognition is independent of attentional focus. *Frontiers in Psychology*, *4*. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2013.00552

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, *63*(6), 902–915. https://doi.org/10.1016/j.neuron.2009.09.006

Öhlschläger, S., & Võ, M. L.-H. (2017). SCEGRAM: An image database for semantic and syntactic inconsistencies in scenes. *Behavior Research Methods*, *49*(5). https://doi.org/10.3758/s13428-016-0820-3

Öhlschläger, S., & Võ, M. L.-H. (2020). Development of scene knowledge: Evidence from explicit and implicit scene knowledge measures. *Journal of Experimental Child Psychology*, *194*, 104782. https://doi.org/10.1016/j.jecp.2019.104782

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. International Journal of Computer Vision, 42(3), 145–175. https://doi.org/10.1023/A:1011139631724

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. Trends in Cognitive Sciences, 11(12), 520–527. https://doi.org/10.1016/j.tics.2007.09.009

Park, J., Josephs, E., & Konkle, T. (2024). Systematic transition from boundary extension to contraction along an object-to-scene continuum. *Journal of Vision*, *24*(1), 9. https://doi.org/10.1167/jov.24.1.9

Peelen, M. V., Berlot, E., & de Lange, F. P. (2024). Predictive processing of scenes and objects. *Nature Reviews Psychology*, *3*(1), Article 1. https://doi.org/10.1038/s44159-023-00254-0

Pellicano, E., & Burr, D. (2012). When the world becomes 'too real': A Bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, *16*(10), 504–510. https://doi.org/10.1016/j.tics.2012.08.009

Perdreau, F., & Cavanagh, P. (2015). Drawing experts have better visual memory while drawing. *Journal of Vision*, *15*(5), 5. https://doi.org/10.1167/15.5.5

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), Article 11. https://doi.org/10.1038/14819

Roberts, K. L., & Humphreys, G. W. (2010). Action relationships concatenate representations of separate objects in the ventral visual system. *NeuroImage*, *52*(4), 1541–1548. https://doi.org/10.1016/j.neuroimage.2010.05.044

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Santi, A., Raposo, A., Frade, S., & Marques, J. F. (2016). Concept typicality responses in the semantic memory network. *Neuropsychologia*, *93*, 167–175. https://doi.org/10.1016/j.neuropsychologia.2016.10.012

Sayim, B., & Cavanagh, P. (2011). What line drawings reveal about the visual brain. *Frontiers in Human Neuroscience*, *5*, 118. https://doi.org/10.3389/fnhum.2011.00118

Schuetz, I., Baltaretu, B. R., & Fiehler, K. (2024). Where was this thing again? Evaluating methods to indicate remembered object positions in virtual reality. *Journal of Vision*, *24*(7), 10. https://doi.org/10.1167/jov.24.7.10

Schwitzgebel, E. (2008). The Unreliability of Naive Introspection. *The Philosophical Review*, *117*(2), 245–273.

Seymour, K., Stein, T., Sanders, L. L. O., Guggenmos, M., Theophil, I., & Sterzer, P. (2013). Altered Contextual Modulation of Primary Visual Cortex Responses in

Schizophrenia. *Neuropsychopharmacology*, *38*(13), 2607–2612. https://doi.org/10.1038/npp.2013.168

Shi, F., Sun, W., Duan, H., Liu, X., Hu, M., Wang, W., & Zhai, G. (2021). Drawing reveals hallmarks of children with autism. *Displays*, *67*, 102000. https://doi.org/10.1016/j.displa.2021.102000

Singer, J. J. D., Cichy, R. M., & Hebart, M. N. (2023). The Spatiotemporal Neural Dynamics of Object Recognition for Natural Images and Line Drawings. *Journal of Neuroscience*, *43*(3), 484–500. https://doi.org/10.1523/JNEUROSCI.1546-22.2022

Spaak, E., Peelen, M. V., & de Lange, F. P. (2022). Scene Context Impairs Perception of Semantically Congruent Objects. *Psychological Science*, *33*(2), 299–313. https://doi.org/10.1177/09567976211032676

Stein, T., Kaiser, D., & Peelen, M. V. (2015). Interobject grouping facilitates visual awareness. *Journal of Vision*, *15*(8), 10. https://doi.org/10.1167/15.8.10

Torralbo, A., Walther, D. B., Chai, B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2013). Good Exemplars of Natural Scene Categories Elicit Clearer Patterns than Bad Exemplars but Not Greater BOLD Activity. *PLOS ONE*, *8*(3), e58594. https://doi.org/10.1371/journal.pone.0058594

Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology*, *59*(11), 1931–1949. https://doi.org/10.1080/17470210500416342

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: Predictive coding in autism. *Psychological Review*, *121*(4), 649–675. https://doi.org/10.1037/a0037665

van den Oever, F., Gorobets, V., Saetrevik, B., Fjeld, M., & Kunz, A. (2022). Comparing Visual Search between Physical Environments and VR. *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 411–416. https://doi.org/10.1109/ISMAR-Adjunct57072.2022.00089

Võ, M. L.-H. (2021). The meaning and structure of scenes. *Vision Research*, *181*, 10–20. https://doi.org/10.1016/j.visres.2020.11.003

Võ, M. L.-H., Boettcher, S. E., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, *29*, 205–210. https://doi.org/10.1016/j.copsyc.2019.03.009

Võ, M. L.-H., & Wolfe, J. M. (2013). Differential Electrophysiological Signatures of Semantic and Syntactic Scene Processing. *Psychological Science*, *24*(9), 1816–1823. https://doi.org/10.1177/0956797613476955

von Helmholtz, H. (1867). *Treatise on Physiological Optics Vol. III* (Bd. 3). Dover

Publications.

Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., & Fei-Fei, L. (2011). Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(23), 9661–9666. https://doi.org/10.1073/pnas.1015666108

Wang, G., Chen, L., Cichy, R. M., & Kaiser, D. (2024). *Enhanced and idiosyncratic neural representations of personally typical scenes* (S. 2024.07.31.605915). bioRxiv; bioRxiv. https://doi.org/10.1101/2024.07.31.605915

Wang, G., Foxwell, M. J., Cichy, R. M., Pitcher, D., & Kaiser, D. (2024). Individual differences in internal models explain idiosyncrasies in scene perception. *Cognition*, *245*, 105723. https://doi.org/10.1016/j.cognition.2024.105723

Wechsler, D. (2009). Wechsler Memory Scale—Fourth Edition. *Pearson*. https://doi.org/10.1037/t15175-000

Wilson, C. J., & Soranzo, A. (2015). The Use of Virtual Reality in Psychology: A Case Study in Visual Perception. *Computational and Mathematical Methods in Medicine*, *2015*(1), 151702. https://doi.org/10.1155/2015/151702

Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, *15*(2), 77–84. https://doi.org/10.1016/j.tics.2010.12.001